bio-protocol Preprint

Not Peer Reviewed
Ask a question online

Gene regulatory network analysis

Liangtao Zheng Shishang Qin Xueda Hu Zemin Zhang *

Updated date: Jan 12, 2022

*For correspondence: zemin@pku.edu.cn

An abbreviated version of this protocol was published in Science in Dec 2021

Pan-cancer single-cell landscape of tumor-infiltrating T cells

DOI: 10.1126/science.abe6474

Detailed protocol

While the datasets from the 10x 5' protocol have a large number of cells, datasets from the SmartSeq2 protocol have higher quality and lower dropout rates. Combining datasets from the two protocols will improve the inference of the gene regulatory network. Therefore, the newly generated 10x data and previously published SmartSeq2 data by us were used for gene regulatory network analysis, as they had the largest number of cells and low batch effects. The $CD8^+$ T cells data and $CD4^+$ T cells data were processed separately, and data from the same protocol were pooled together. To reduce the dropout impact on gene-gene relationship inference, the analysis was based on mini-clusters. The gene expression profiles of the mini-clusters were fed into the python implementation of the SCENIC algorithm (pyscenic). The *grn* command of pyscenic was used to infer the gene-gene co-expression relationships between transcription factors (TF) and their potential targets, while the TF list containing genes annotated as "transcription factor activity" by GO (GO:0003700) was used, and "grnboost2" was selected as the gene regulatory network reconstruction algorithm. After each run of *grn*, it issued a file containing a list of the network's adjacent edges with TF name, potential target gene, and an associated importance score. To combine the data from the two protocols, for the same adjacent edge, the median importance scores from the data of the two protocols were used in the combined result. The combined adjacent edge list was further filtered to keep only the TF-target pairs with an importance score > 1. Then the adjacent edge list was used as input of command *ctx* to identify the regulons each of which contained one TF and its target genes enriched for the motifs of the TF. The motif annotation database (--annotations_fname option: motifs-v9-nr.hgnc-m0.001-o0.0.tbl) and the ranking databases (database_fname parameter: hg19-500bp-upstream-7species.mc9nr hg19-tss-centered-10kb-7species.mc9nr) used by the *ctx* command were downloaded from https://resources.aertslab.org/cistarget. The *ctx* command was run for the data from the two protocols respectively. After that, command *aucell* was used to calculate the activities of the regulons for each mini-cluster. Given the activity scores of the regulons and the meta-cluster assignment of the mini-clusters, python function *regulon_specificity_scores* was used to calculate the specificity scores of the regulons (RSS). At last, for each regulon, the median of the RSS from data of the two protocols was used to rank the regulons.
The code could be found in Zenodo (10.5281/zenodo.5461803), scripts under the folder code/scenic.

How to cite:(Readers should cite both the Bio-protocol preprint and the original research article where this protocol was used)

1. Zheng, L. , Qin, S. , Hu, X. and Zhang, Z. (2022). Gene regulatory network analysis. Bio-protocol Preprint. bio-protocol.org/prep1502.
2. Zheng, L., Qin, S., Si, W., Wang, A., Xing, B., Gao, R., Ren, X., Wang, L., Wu, X., Zhang, J., Wu, N., Zhang, N., Zheng, H., Ouyang, H., Chen, K., Bu, Z., Hu, X., Ji, J. and Zhang, Z.(2021). Pan-cancer single-cell landscape of tumor-infiltrating T cells. Science 374(6574). DOI: 10.1126/science.abe6474